Evaluation for Collaborative Systems

Laurie Damianos, Lynette Hirschman, Robyn Kozierok, Jeffrey Kurtz,
The MITRE Corporation
202 Burlington Road
Bedford, MA 01730 USA
+1 781 271-2000
(laurie, lynette, robyn, jkurtz) @mitre.org

Andrew Greenberg, Ruel Holgado, Kimberley Walls
NIMA
NIMA Washington Navy Yard, 1st and M SE,
Washington, D.C. 20505
+1 202-314-5548
(greenberga, HolgadoR)@nima.mil
kimw@ucia.gov

Sharon Laskowski, Jean Scholtz
NIST
Bldg. 225, A216
Gaithersburg, MD 20899
+1 301-975-4535
(sharon.laskowski, jean.scholtz)@nist.gov

ABSTRACT

The Evaluation Working Group (EWG) in the Defense Advanced Research Projects Agency (DARPA) Intelligent Collaboration and Visualization (IC&V) program has developed a methodology for evaluating collaborative systems. This methodology consists of a framework for classification of CSCW (Computer Supported Cooperative Work) systems, metrics and measures related to the various components in the framework, and a scenario-based evaluation approach. This paper describes the components of this methodology. Two case studies of evaluations based on this methodology are also described.

Keywords

Evaluation, computer supported cooperative work, scenario-based evaluation, case studies, metrics, measures.

INTRODUCTION

The Evaluation Working Group (EWG) of the Intelligent Collaboration and Visualization (IC&V) program was established to develop the metrics and evaluation methodology for the collaborative technologies that make up the IC&V program. Additionally, the EWG was to develop, or guide the development of, specific tests and tools for achieving effective and economical evaluation. Evaluation is an important tool for all researchers and developers because it allows classifications and comparisons to be made. When used in iterative software development, successive evaluations allow us to measure progress. Evaluation also permits us to determine if a system meets particular user needs, or if it provides an improvement over existing systems. The evaluation methodology for CSCW systems developed by the EWG is not specific to the IC&V program, but can be used by any group developing or using CSCW systems.

The EWG[1] has taken as its primary task the definition and validation of low-cost methods of evaluating collaborative environments, such that researchers in the collaborative computing research community can use these methods to evaluate their own or other research products. The group's Evaluation Methodology Document [2] describes a strategy designed to meet the following goals:

- To develop, evaluate and validate metrics and a methodology for evaluating collaborative tools.
- To provide reusable evaluation technology, such that research groups can assess their own progress.
- To provide evaluation methods that are inexpensive relative to the requirements.
- To apply Department-of-Defense-relevant criteria.
- To define an application vision that will drive collaborative computing research.

The methodology developed by the EWG consists of a framework for describing CSCW systems, metrics for evaluating the various components of a CSCW system, and

[1] The EWG comprises researchers from several sites with diverse backgrounds and interests. The organizations represented are Carnegie Mellon University (CMU), the MITRE Corporation, the National Imagery and Mapping Agency (NIMA) and the National Institute of Standards and Technology (NIST).

a scenario-based evaluation technique. In this paper we will describe this framework and the metrics. We will also explain how scenario evaluation works. In the last two sections we present case studies using this evaluation methodology.

EVALUATION OF COLLABORATIVE SYSTEMS

The success of evaluations is dependent on defining the appropriate hypotheses or objectives, selecting the relevant data collection techniques, and conducting the evaluation in an objective and repeatable fashion. The EWG Methodology document provides guidance for designing and conducting evaluations that adhere to these principles. Our approach to evaluation consists of the following steps:

- Formulate hypotheses
- Determine appropriate evaluation method(s) and scope
- Select data collection instruments and identify measures

Each of these is described in detail below.

Formulate Hypotheses

Hypotheses are specifically stated predictions about the relationships among independent and dependent variables. They are the driving factor for determining all other aspects of conducting an evaluation.

Determine Appropriate Evaluation Method(s) and Scope

The selection of methods for evaluating the utility of collaborative tools, and systems in general, is dependent on the proposed scope of the evaluation, including (1) the specificity of the research question, (2) system maturity, (3) control of the experimental variables desired, (4) availability of representative participants, (5) level of confidence needed in the results, (6) time constraints, (7) resource requirements, and (8) the usability issues being addressed [4]. Performing this cost-benefit tradeoff between instances of these factors with the users of the evaluation better ensures that the results are useful. The Evaluation Toolkit [4] provides a structure for walking through these alternatives with a user. The toolkit examines the applicability of various data collection techniques based on the answers to questions associated with the factors just cited.

Another approach to selecting an evaluation method depends upon one's evaluation goals. Various evaluation questions can be addressed using different evaluation methods. Different types of evaluation and the associated questions addressed are shown in Table 1.

Once we have determined our evaluation goal, we can choose an evaluation method and determine the resources required to carry it out. For example, an appropriateness evaluation can be accomplished by comparing the group's functionality requirements to the capabilities existing in the system. A comparative or iterative evaluation is best carried out using appropriate scenarios with representative users in a lab experiment or, if the product is robust enough, via a field study with users in their environment, involved in the performance of mission-driven tasks.

Question	Type of Evaluation
Does the system run, and can we afford it?	Feasibility evaluation
Have improvements been made to a system?	Iterative evaluation
Is system A better than system B?	Comparative evaluation
Does the system have the necessary capabilities?	Appropriateness evaluation

Table 1: Evaluation Questions

Select Data Collection Instruments and Identify Measures

Data collection instruments are the means of taking both direct measurements and indirect measurements. Direct measurements involve measuring individual actions, such as start and end times or the number of steps a user must perform to reach a desired end state. Indirect measurements involve making inferences from the responses to a questionnaire or from measurements taken.

Data from the evaluations can be collected in numerous ways, including:

- Logging tools for collecting a time-stamped record of participant actions
- Direct observation
- Questionnaires/interviews/rating scales (open-ended or closed/fixed alternatives)
- Video and audio recordings

In selecting the method to use for data collection, care should be taken to make sure that data collection does not disrupt system performance or user-system interaction. For example, a logging tool might slow system response time, which might adversely affect the performance of user tasks.

Issues in Evaluating Collaborative Systems

Collaboration involves multiple humans interacting with networked systems, making the evaluation at least an order of magnitude more complex than typical HCI (human-computer interaction) evaluations. The heuristic or expert reviews used effectively for single user interfaces may not take into account these additional social interactions. In addition, evaluating technical performance across multiple networked environments is inherently more difficult than stand alone assessments.

Evaluation design for collaborative applications and environments is thus more difficult. Helping researchers determine the appropriate items to track and measure and providing guidance on evaluation methods facilitates evaluation, which is an important element in iterative software development. This feedback provides input all along the development cycle, allowing a more rapid deployment of useful technology.

Collaborative systems are diverse. They are used by many different groups for many different purposes. In order to guide researchers in selecting the appropriate evaluation techniques and to help in interpreting the evaluation results, a structure is necessary. The next section describes a framework developed by the EWG for classifying collaborative systems.

EWG FRAMEWORK FOR COLLABORATIVE SYSTEMS

The goal of the framework is to facilitate description of a collaborative system and to evaluate how well that system supports various kinds of collaborative work. Users should be able to utilize our framework top-down to map their requirements to CSCW technologies and bottom-up to ascertain how well a given CSCW system supports their work. Developers of CSCW systems should be able to use our framework to describe their system and to determine the types of group work that could be easily accomplished using the system.

This framework builds on one devised by Pinsonneault and Kraemer [10] to analyze the impact of technology on group process while controlling for the effect of other contextual variables. We have merged the work of McGrath [6] into our expanded framework to enable us to classify tasks that groups perform. As shown in Figure 1, the framework is divided into four levels: requirement, capability, service, and technology.

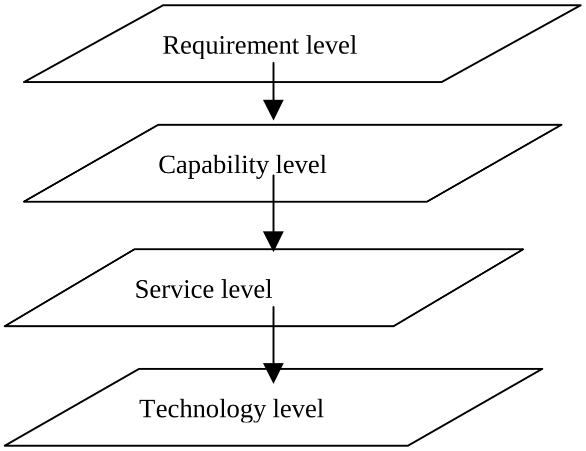

Figure 1: The 4 level framework

Requirement Level

The requirement level describes the requirements of the group with respect to the tasks being performed and the support necessitated by the characteristics and social protocols of the group. It is divided into four sections: work tasks, transition tasks, social protocol requirements, and group characteristics.

Work Tasks

The following collaborative work tasks are defined in the framework:

- Planning
- Brainstorming and group creativity
- Intellective
- Decision making
- Cognitive conflict
- Mixed motive tasks, including bargaining and negotiation
- Competitive performance
- Non-competitive performance
- Information dissemination

For each work task defined in the framework, specific measures are given that are appropriate to use in evaluating the outcome of the task and the process of doing the task. Where possible, we have used existing research to suggest capabilities that should be useful in doing this task. As an example, Figure 2 shows the framework description for a brainstorming task type.

Transition tasks

Transition tasks are tasks used to move between work tasks in synchronous or asynchronous collaborations. These tasks include such things as summarizing the outcome of the task just completed or assigning action items to members of the group. Additionally, meeting setup tasks are needed at the start of synchronous collaborations.

Social Protocols

Social protocols define the ways in which collaborative sessions are conducted. Collaborative sessions may vary from informal sessions to very formal sessions with a chair, an agenda that is strictly followed, and rules of order. Social protocol capabilities support communication between group members; awareness of group members, group activities, group objects; and basic meeting conduct.

Group Characteristics

Groups have different requirements depending on the makeup of the group, the location of the group members and the time requirements for collaborative sessions. We classified the dimensions as time related, group type, and computer related. Time dimensions refer to the length of sessions, the overall length of the collaborative effort, and whether sessions are spontaneous or planned. Dimensions of group type consider the number of people in the group, whether the group is newly formed or has been working together for some time, how homogeneous members of the group are, whether the groups are geographically distributed, and whether the groups must work synchronously. Computer related dimensions include the hardware platforms of the group, the time available for training, and the computer expertise of group members.

Group characteristics can change over time. For example, a group could initially consist of five members but know that, in the space of several months, another ten members will be added. This should be taken into account during the specification of requirements, if possible.

Task: Brainstorming and group creativity

Members of a group are given a particular topic area and asked to brainstorm on ideas.

Specific measures for this type of task:

- Number of ideas
- Originality of ideas

Known research:

- Creativity of individuals is stifled by social influence of group
- Individuals are able to take advantage of creativity-enhancing forces in group - social support, cross stimulation

Example: The group has a goal to raise $200,000 to build a new community center. They generate ideas for funding raising events, people to ask for contributions and possibilities for loans or selling "shares" to the community members to raise this money.

Suggested capabilities:

- Anonymous communication
- Synchronous communication
- N way communication
- Shared workspace

Figure 2: Example Work Task Description

Capability Level

The capability level of the framework describes functionality that is needed to support the requirements. This includes such things as the ability to share documents and programs, support for awareness of other people and activities in the workspace, communication, etc. Capabilities support work tasks, transition tasks, and social protocols needed by the groups. Capabilities supporting work tasks include object manipulation, visualization of objects, and object management. Transition task capabilities include collaborator locator support, agenda support, calendar support, voting, playback, and summarization. Social protocol capabilities include indicators of awareness, floor control mechanisms, synchronization features, side chat capabilities, anonymous communication, and private workspaces. Capabilities supporting group requirements include the number of participants that must be supported and the need for synchronous or asynchronous collaborations.

Service Level

The service level describes specific services that can be used to provide the capabilities needed for collaboration. An example list of services that could be combined to provide CSCW capabilities include e-mail, a chat facility, telephone connections, multicast audio, multicast video, a white board, shared applications, encryption, a history mechanism, version control, collaborative space navigation, and collaborative space management. Different services can be used to provide the same capability. For example: the capability to have a side chat with another meeting participant during an electronic meeting could be accomplished by a text chat service or by telephone service.

Technology Level

The technology level describes specific implementations of services. This level could be viewed as the set of all possible components needed to build a given collaborative system, including integration and user interface components.

MEASURES AND METRICS FOR EVALUATION

The goal of the Evaluation Working Group is to define inexpensive evaluation technology that can be reproduced in a laboratory. Hence, we have not provided measures to address organizational impact and other measures that require long term field studies of systems.

Experiments involving human subjects, or sets of subjects, are expensive and time consuming in terms of obtaining participant time for studies, data collection and analysis. In many cases, the measures must be developed and validated before they can be applied with confidence. Despite these difficulties, we lay out here some options for evaluating collaborative systems at all four levels of the framework. The following are examples of measures that are applicable at the various levels:

Requirement level: task outcome, user satisfaction, efficiency, scalability, degree of security.

Capability level: collaboration management, transition support, collaborative object support, task focus.

Service level: quality of service, audio, video, and image quality.

Technology level: usability measures, network load.

Metrics are indicators of system, user, and group performance that can be observed, singly or collectively, while participants perform tasks in an operational environment or while they execute actions defined in a scenario. Metrics - like time, length of turn, and other countable events - are directly measurable and can be collected automatically.

Measures can be taken at each of the four levels of the collaborative framework. For example, task completion time (a requirement-level measure) is based on start and end time metrics. A measure can also be a combination of interpreted metrics and other measures. A complicated measure, like efficiency, is partially derived from the interpretation of metrics like time, user ratings, and tool usage. In addition, measures of system breakdown (taken at the service level) contribute to efficiency. A simple way to distinguish between metrics and measures is to apply the following criterion: a metric is an observable value, while a measure associates meaning to that value.

The EWG Methodology document defines measures that could be used for the evaluation of CSCW systems. Along with each measure we give a definition, the ``building blocks" --metrics and other measures, and the associated task types. For example, at the capability level the measures for collaboration management are defined as shown in Figure 3.

Collaboration Management Measures

The set of collaboration management measures assesses support for coordinating collaboration.

These measures are used to evaluate the following (non-exhaustive) list of capabilities:

- support for multiple collaborations
- floor control
- agenda support
- document access control
- collaborator access control
- synchronize feature

Aspects of group work evaluated

- work task
- transition
- social protocol

Metric and measure components

- expert judgments - yes/no, to what degree

Additional metrics and measures that may be relevant or correlated

- turn overlap

Associated task types

- all

~~Figure 3: An example of a definition of a set of measures~~

SCENARIOS FOR EVALUATION

One versatile tool for an evaluator's repertoire is scenario-based evaluation. Scenarios can be used for many development activities including design, evaluation, demonstration, and robustness testing. When used for evaluation, scenarios exercise a set of system features or capabilities in a manner consistent with the tool's operational use. Bass [1] identified scenarios that could be used to design, evaluate, illustrate, and compare CSCW systems. Nardi [9] argued that a library of reusable, salient scenarios should be created. Potts [11] emphasized the importance of identifying goals and obstacles to produce relevant scenarios.

The EWG has provided a suite of scenarios for evaluation (http://www.antd.nist.gov/~icv-ewg/pages/scenarios.html). In addition, our Methodology document provides guidance for researchers and developers who wish to construct their own scenarios for use in evaluation. In terms of the collaborative framework, a scenario is an instantiation of a generic work task type (as defined at the requirement level of the framework), or a series of generic work tasks linked by transitions. It describes what the users are expected to do, such as analyze imagery or create a logistical plan in a given context. It usually specifies the characteristics of the group carrying it out, and the social protocols used by that group.

The various scenarios provided by the EWG emphasize different work tasks and different group characteristics. Researchers and others wishing to conduct evaluations may be able to select a scenario that matches their group requirements. The EWG encourages contributions of scenarios that others may have developed for evaluation purposes. Once a scenario is constructed, it can be reused to evaluate new versions of a collaborative system or other collaborative systems. Reuse of that scenario, complete with measures and metrics of interest to the group, fosters repeatability.

The following sections describe two case studies: one conducted primarily to validate the evaluation methodology and a second case study where the methodology was used to evaluate the appropriateness of a collaborative tool for use by a specific group.

CASE STUDY: MITRE MAP NAVIGATION EXPERIMENT

This case study describes how the Evaluation Methodology was used to evaluate a collaborative computing environment at the MITRE Corporation. Our primary goal was to apply and validate our methodology. A secondary goal was to test the MITRE Multi-Modal Logger [6] in collecting time stamped data. This report focuses on the role of the EWG Methodology in guiding our evaluation. The Map Navigation Report [5] contains details of this evaluation.

Tasks and the Scenario

We designed a simple experiment to see how communication modalities would influence planning and the sharing of information. The experiment compared two configurations of a collaborative environment: one configuration supported audio and text conferencing and one supported just text conferencing. The design process involved choosing a scenario based on the prototypical kinds of collaborative work tasks outlined in the EWG Methodology document (e.g., decision making, planning, information dissemination, etc.). We formulated hypotheses about the influences of the communication modality. We also considered practical issues of data collection in a laboratory environment, particularly finding willing subjects knowledgeable enough about an artificial task to carry it out. Research topics for collaborative environments, such as awareness, grounding, and modality coordination, directed our focus as well.

We constructed a scenario that involved two types of work tasks: a planning task and an information dissemination task. The scenario was collaborative map navigation. A pair of participants, who communicated via a collaborative environment, shared an on-line street map. Each was given private information, about congestion and traffic restrictions, that had to be shared in order to plan a valid route between two points on the map. This scenario was chosen over a similar, military scenario that involved troop movement and obstacle avoidance because our pool of participants was more familiar with road map navigation.

The social protocol for the scenario was that of an informal session between two anonymous co-workers[2]. The environment consisted of MITRE's Collaborative Virtual Workspace, [8, 13] which included a shared whiteboard, text conferencing and audio conferencing. When introducing the audio tool to the participants, the experimenters emphasized that only one participant could talk at a time[3].

Hypotheses and Measures

Once we established a scenario that highlighted our research interests, we formulated hypotheses and consulted the Methodology document for appropriate measurements. The hypotheses focused on a comparison of audio and text-only modalities, as follows:

- People would collaboratively plan a route faster when audio communication was available.
- People would collaboratively plan a better route when audio communication was available.
- Participants would be more satisfied with collaborative route planning when audio communication was available.
- Participation would be more equal when audio communication was not available.
- The number of speech turns when audio was used would be greater than the number of typed turns in the non-audio condition.
- The whiteboard would be used less in audio mode than in non-audio mode.

To test these hypotheses, we designed the experiment to have eight pairs of participants perform the scenario under both system configurations: with audio conferencing and with text only.

Next we identified exactly what to measure. The Methodology document was helpful in describing the appropriate metrics to collect at each level of the framework.

At the requirement level, we evaluated how well our collaborative system supported the scenario and the group characteristics. At this level, we also evaluated the artifacts produced (the final route) as well as the strategy the group used in producing the final route. We measured group process by time to reach a decision, efficiency of communication, and equality of participation.

At the capability level, we were interested in evaluating how well given capabilities, such as shared workspace and two-way communication, supported collaborative planning and information sharing. To obtain answers to these questions, we used measures such as task completion time.

Although we did not plan to evaluate the system at the service level, we did get some feedback on the difficulties of using certain services simultaneously, namely text chat and whiteboard.

We also did not choose to evaluate the system at the technology level, but participants did comment on usability issues and acceptability of the tools. Most complaints were about the difficulties relating to usage of the audio tool (configured for half-duplex): not being able to hear the other person when participants were speaking simultaneously. This impeded gaining floor control. In addition, the whiteboard lacked a pointing device which made it difficult to draw the other participant's attention to newly-added information.

Below we detail the metrics of the requirement and capability levels, based on our hypotheses.

Requirement Level Metrics and Measures

Task outcome measures included whether or not the task was completed and the quality of the final route. The route score was determined by a formula, which factored in distance, road speed, traffic conditions, and construction delays. Scores were penalized where violations occurred, like driving through a roadblock.

Overall task completion time was taken as the time difference between the start of the scenario and the ending

[2] We wanted to keep the identities of each participant anonymous because we thought familiarity with one another might influence their behavior and interactions. In particular, we thought a person's professional status (or even age) might affect the other participant's participation. As a partial enforcement of anonymity, the participants were never in the same room during any part of the experiment. We also used code names for their character representations in CVW.

[3] We configured the audio tool for half-duplex so that observers of the experiment could listen in on the audio via speakers without getting feedback.

transition task where the participants had agreed on the final route.

User satisfaction was measured by user ratings, as established by a questionnaire.

Participation measures included the number of words per participant, the number of turns per participant, the time spent communicating in each mode, and various user ratings.

Consensus was a necessary condition to the completion of the task but was checked by comparing each participant's version of the final solution.

Capability Level Metrics and Measures

Communication measures included the number of turns per participant, user ratings (on the goodness of communication, the ability to get floor control, the ability to get the attention of the other participant, and the ability to interrupt), and expert ratings (on the goodness of communication and the ability to get floor control).

We defined three types of turns: typing, speech, and whiteboard. A typing turn consisted of a text message and a carriage return; a speech turn was logged when silence was detected after an utterance; and a whiteboard turn was a single annotation.

We also collected the time each participant spent in each communication modality: time spent typing, time spent speaking, and time spent drawing on the whiteboard.

Data Collection

We used observations, questionnaires and logs to gather the measures and metrics. Experimenters recorded the participants' comments during and after the task to gain insight into service and technology level issues. The questionnaires assessed the participants' satisfaction and perceived participation.

The MITRE Multi-Modal Logger was used to record time stamped speech, typed text, whiteboard activity, and other user events. The time metrics detailed above were obtained from the time stamped events.

Results and Observations

The results of our analysis show that task outcome was affected by the groupware configuration used; performance was significantly faster but not significantly better when the participants used audio to communicate. Satisfaction was nominally the same in both conditions, but a preference was observed for the audio condition. Level of participation between conditions was not significant. Finally, a significant difference in the usage of communication utilities was seen between the conditions. For example, the whiteboard was used more when audio was not available.

In addition, we uncovered some interesting design implications for collaborative systems where audio may not be available. Typing and whiteboard events appeared in different windows; users sometimes were not aware of one type of event when they were focused on another. It required some explicit effort for one participant to draw the other's attention to a new whiteboard annotation, leading to miscommunication and inefficiencies. The system would benefit from better support for multi-modal awareness.

Lessons Learned

The Methodology document served two key purposes in our comparison evaluation; it assisted in identifying a scenario based on the work tasks and also provided us with relevant measures and metrics. We found the scenario a good tool for evaluation, and users enjoyed participating in the experiment. We believe we obtained useful data.

Although the listed measures and metrics were helpful as suggestions for what to measure, we wanted more guidance on how to take the actual measures. Metrics that originally seemed easy to measure, like time to complete task and the number of whiteboard turns, proved to be more complicated than we thought. It was difficult to create a logged event associated with the start times of both participants because they were not co-located. Regarding whiteboard turns, there are several modes of drawing in the whiteboard (i.e., line segments vs. a continuous curve) that look different when logged. Line segments are logged as multiple events whereas a long curve is logged as a single event, but both the line segments and the long curve really represent a single 'turn'.

During the experiment, observers noticed indications of awareness and focus issues. The audio condition allowed participants to talk while focusing on other activities; without audio, participants could not type and focus on multiple activities at the same time. On several occasions, one participant appeared to be unaware of what the other participant was doing. In addition to these observations, interviews with the users provided some insight, but a more methodical approach was lacking. The research community needs a clear method for measuring awareness and focus.

A section on experimental design should be included in the Methodology document. This section should emphasize the importance of using pilot studies during the development of the experiment. Our pilot studies were extremely useful in elucidating flaws in our experiment as well as helping us determine if the data we were collecting were adequate. The studies assisted in refining our hypotheses and modifying our data collection to include additional metrics.

In conclusion, we liked the scenario approach to evaluating systems. Particularly useful were the suggested measures for each particular work task. The hypotheses helped us select from the set of measures and metrics identified in the Methodology document. We would like to recommend adding a section on experimental design as well as elaborated details on how to gather the particular metrics and measures.

CASE STUDY: NIMA EVALUATION OF PLACEWARE™

Introduction

The purpose of the NIMA evaluation of PlaceWare[4] was to document the utility of PlaceWare for supporting collaborations in an operational setting. As the paradigm for PlaceWare's use reflects the one-to-many dissemination of information in an auditorium setting, the evaluation of this tool examined PlaceWare's utility at supporting meetings within NIMA. Framework requirement level tasks addressed in this evaluation include "information dissemination" and "planning." These tasks involved:

- Supporting a subset of division personnel (a group of 20 to 30 personnel) who can not attend a scheduled meeting because they are separated geographically.
- Supporting a small group of personnel separated geographically, with a need to communicate and share information toward a common objective. This includes one-to-one collaborations and collaborations involving three and four participants.

Although PlaceWare can support more than the four to five participants we used in any one session, the infrastructure (being connected through a sensitive but unclassified (SBU) Ethernet network) and hardware limitations within our organization prevented us from evaluating this tool under conditions involving more personnel. In this instance, although we used the framework to determine group requirements, we were limited in our ability to evaluate them.

The Experimental Design

Evaluation Objectives

The objectives for this operational evaluation were to:

- Document the adequacy of PlaceWare capabilities for supporting small group meetings within NIMA.
- Document the usability of PlaceWare capabilities.
- Identify other potential uses of the PlaceWare application.
- Identify gaps in the collaborative computing technology being evaluated as well as this technology in general for its ability to meet the unique requirements of the participants.

Participants

The participants in this study included scientists and engineers and their managers within technology domains. These personnel participated from three geographically separate locations.

Network Environment

PlaceWare was evaluated over multiple platforms and networks. During pre-testing, PlaceWare was used (a) over the Internet, with access provided by Internet Service Providers (ISPs), (b) on a closed testbed LAN, and (c) using Pentium-compatible computers residing on a sensitive but unclassified (SBU) 10 MBPS Ethernet network.

Methods

We collected data through:

- Post collaboration user interviews
- Evaluator observations.
- A review of capability adequacy and ease of use by NIMA human computer interface and systems personnel.

Results

A detailed description of the PlaceWare evaluation may be found in Walls, K., Greenberg, A., and Holgado, R. [14]. PlaceWare capabilities support one-on-one, one-to-many, and many-to-many collaboration requirements. The web browser-based graphical user interface is consistent with current user interface technology and is very easy to use. Overall, PlaceWare supported group meetings very well.

The auditorium paradigm worked well at supporting different moderator styles. One moderator preferred that remote attendees only communicate at specific times and thus liked the PlaceWare attendee hand raising and question capabilities; whereas, another moderator liked the free flow exchange of information, and thus had remote participants also log in as presenters. This demonstrates PlaceWare's flexibility in terms of the framework's social protocol dimension. Another result that relates to the social protocol dimension of the framework is PlaceWare's inability to easily support the identification of all collaboration participants. At present, only the attendees within a single auditorium row are listed at any one time.

PlaceWare had two primary collaboration features: audio / chat communications and a slide presentation tool.

Audio / Chat

The limited available bandwidth via the dial-up connections over the Internet generated considerable performance degradation during our tests. Primary areas of degradation include slow system performance with login procedures and audio distortion and dropout. Under standard 10 MBPS Ethernet conditions, overall audio is good in terms of dropout, distortion, background noise, and delay.[5] Text chat was identified as a good complement to audio, and was especially useful for (1) conducting one-on-one side exchanges during a presentation and (2) recording key points and action items when used with the logging capabilities.

[4] PlaceWare is a collaborative tool supporting Windows NT/95, HP-UX, and Unix with audio, chat, slide presentation, logging, and voting capabilities.

[5] During one collaboration, however, there was a significant amount of distortion and dropout due to heavy network traffic and the use of microphones not suited for the characteristics of the meeting.

Presentation Tool

Evaluation participants and evaluators indicated that although presentation capabilities are adequate for presenting ideas, two problems surfaced: (1) Slide creation capabilities need to be interoperable with other applications, in addition to Power Point, and (2) slide presentation font requirements are too restrictive (i.e. users are required to use a minimum of 20 size font).

In addition to examining the utility of PlaceWare for small group exchanges of information, we attempted to use PlaceWare to support a brainstorming activity, realizing that it was not an intended use. We found that PlaceWare did not support this activity well because:

- The viewable information jumped as multiple people typed in new data or as they scrolled the screen to see prior inputs.
- There was no word wrap feature in the presenter tool window, which made it difficult to view comments easily.
- Sometimes individuals would accidentally type into the same workspace as another participant.

Lessons Learned

PlaceWare capabilities supported NIMA small group collaboration needs well.

One major lesson learned is that appropriate system configuration is critical and much care should be taken when selecting peripheral devices. Microphones incorporating echo/noise cancellation techniques may have alleviated some audio problems when remote participants collaborated with a full room of participants.

This evaluation illustrates the application of many of the EWG Framework concepts. For example, prior to integrating PlaceWare into the operational environment, we examined PlaceWare at the requirement and capability levels. At the requirement level, we determined that audio would be needed to share running meeting comments with remote participants. In addition, the ability to present briefing material is essential in many NIMA meetings. At the capability level, we determined that PlaceWare had the needed functionality required for supporting meeting information exchange requirements. At the service and implementation levels, we determined that PlaceWare, as implemented, addressed our specific requirements. For example, during pilot testing, we determined the minimum size font and font types that would be discernible, once briefing material was fitted within the PlaceWare slide presentation tool. In addition, we examined the PlaceWare user interface attributes through the use of a pre-defined set of user interface questions developed from Human-Computer Interface guidelines, defined by Sebrechts [12].

CONCLUSIONS

The two case studies described in this paper have validated the basic structure of our evaluation methodology. The two case studies were quite different: the MITRE case study was used to compare two versions of a single system with different capabilities with respect to a particular type of task. The NIMA case study was used to evaluate an existing CSCW product with respect to the needs of a particular group. In both instances, the framework was useful in specifying the requirements, selecting the scenarios, and determining the metrics and measures.

We have also obtained some specific information that should be incorporated into our Methodology document. The MITRE study shows the need for a section in the document describing experimental design procedures for researchers unfamiliar with such methods. The study also showed the need for more guidance in taking measures.

Usability issues were encountered in both case studies. These cannot be ignored because serious usability problems can affect how and when participants use different services in the systems. The Methodology document could be augmented with a usability questionnaire. A section on interpreting evaluation results should include how to balance usability issues, capability issues, and overall group performance issues.

The case study conducted by NIMA also pointed out the need for some capability level evaluations that they did prior to the scenario-based evaluation. We are currently working on making these evaluations and similar evaluations available as a resource. This would allow researchers access to some off-the-shelf evaluations for capability tests, thus saving time and facilitating repeatability.

The NIMA case study also called attention to the need to evaluate peripheral devices when configuring the system to be used. The capability level of the framework could be extended to include considerations of capabilities needed in peripheral devices.

We recognize the need for long term ethnographic studies of CSCW systems to determine group and organizational effects [3]. However, we feel that the evaluation methodology we have developed is useful in determining the initial benefits of CSCW systems and in comparing systems with different capabilities. We have identified some areas in our methodology that need augmenting or improving, but the case studies have validated the use of the methodology in providing a low cost, reusable evaluation method.

ACKNOWLEDGMENTS

MITRE's contribution to this work was funded, in part, by DARPA under contract number DAAB07-97-C-C601. Contributions to this work by NIMA and NIST were funded, in part, through the DARPA IC&V Program.

REFERENCES

1.Bass, L. (1996) Mini-workshop: scenarios for CSCW systems. In editors (Bass, L. J. and Unger, C) *Engineering for Human-Computer Interaction: proceedings of the IFIP TC2/WG2.7 working conference on engineering for human-computer interaction*, Yellowstone Park, USA, August 1995, Chapman & Hall, pp. 333-338.

2.Cugini, J, Damianos, L., Hirschman, L., Kozierok, R, Kurtz, J., Laskowski, S, Scholtz, J. (1997). Methodology for the Evaluation of Collaboration Systems, (http://www.antd.nist.gov/~icv-ewg/documents/meth_index.html)

3.Grudin, J. (1988), Why CSCW Applications Fail: Problems in the Design and Evaluation of Organizational Interfaces, in *Proceedings of ACM CSCW'88 Conference on Computer-Supported Cooperative Work, Perspectives on Evaluation*, pp. 85-93.

4.Hall, T., Walls, K. and Monago, A. (1996). *Evaluation Design Toolkit*, NPIC/NIMA Document

5. Kurtz, J, Damianos, L, Hirschman, L, and Kozierok, R. (1997), The Map navigation Experiment, http://www.antd.nist.gov/~icv-ewg/experiments/experiments.html.

6.McGrath, J. E. (1984*), Groups: Interaction and Performance*, Englewood Cliffs, N. J., Prentice-Hall.

7.The MITRE Multi-Modal Logger: http://www.mitre.org/research/logger

8.MITRE's Collaborative Virtual Workspace, http://www.mitre.org/resources/centers/advanced_info/g04e/cvw.html

9.Nardi, B. A. (1995) Some reflections on scenarios. In ed (J. Carroll*) Scenario Based Design: Envisioning Work and Technology in System Development*,. pp. 397-399.

10.Pinsonneault, A. and Kraemer, K. (1993), The Impact of Technological Support on Groups: An Assessment of the Empirical Research in ed (R. Baecker,) *Readings in Groupware and Computer Supported Cooperative Work*, pp. 754-773.

11.Potts, C. (1995*) Using schematic scenarios to understand user needs*, in DIS95, Ann Arbor, MI, pp. 247-256.

12.Sebrechts, M., Knott, B., and Etgen, M. (1997*). Human-Computer Interaction Guidelines for Imagery Exploitation Systems*, Volume 2.0 NIMA Document,.

13.Spellman, P. J. and Carlson, J. Technology for Virtual Organizations. Extended abstract in Proceedings GroupWare '95, Boston, MA, March, 1995.

14.Walls, K., Greenberg, A., and Holgado, R. (1998). National Imagery and Mapping Agency (NIMA) Evaluation of the PlaceWare Auditorium Collaboration Tool, NIMA Document.